

A Semantic Framework for Web-based Accommodation Information Integration



Kai Yang

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2012

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

Abstract

With the tremendous growth of the Web, a broad spectrum of accommodation information is to be found on the Internet. In order to adequately support information users in collecting and sharing information online, it is important to create an effective information integration solution, and to provide integrated access to the vast numbers of online information sources. In addition to the problem of distributed information sources, information users also need to cope with the heterogeneous nature of the online information sources, where individual information sources are stored and presented following their own structures and formats. In this thesis, we explore some of the challenges in the field of information integration, and propose solutions to some of the arising challenges. We focus on the utilization of ontology for integrating heterogeneous, structured and semi-structured information sources, where instance level data are stored and presented according to meta-data level schemas. In particular, we looked at XML-based data that is stored according to XML schemas.

In a first step towards a large-scale information integration solution, we propose a semantic integration framework. The proposed framework solves the problem of information integration on three levels: the data level, process level and architecture level. On the data level,

we leverage the benefit of ontology, and use ontology as a mediator for enabling semantic interoperability among heterogeneous data sources. On the process level, we alter the process of information integration, and propose a three step integration process named as the publish-combine-use mechanism. The primary goal is to distribute the efforts of collecting and integrating information sources to various types of end users. In the proposed approach, information providers have more control over their own data sources, as data sources are able to join and leave the information sharing network according to their own preferences. On the architecture level, we combine the flexibility offered by the emerging distributed P2P approach with the query processing capability provided by the centralized approach. The joint architecture is similar to the structure of the online accommodation industry.

This thesis also demonstrates the practical applicability of the proposed semantic integration framework by implementing a prototype system. The prototype system named the “accommodation hub” is specifically developed for integrating online accommodation information in the large, distributed, heterogeneous online environment. The proposed semantic integration solution and the implemented prototype system are evaluated to provide a measure of the system performance and usage. Results show that the proposed solution delivers better performance with respect to some of the evaluation criteria than some related approaches in information integration.

Acknowledgements

This thesis is fully supported by the ARC Linkage Project, a joint effort between the University of Technology, Sydney and the Lido Group. Here, I would like to acknowledge that this research would not have been possible without the support from both the university and the industry partner.

First of all, I want to thank my supervisor, Prof. Robert Steele, who has brought me to this exciting research project, and guided me throughout my candidature. I would like to express my gratitude to him for providing me with the inspirational ideas and comments for this research.

I also wish to thank all my colleagues from the university of technology, Sydney for the invaluable inputs to this research and the enjoyable teamwork. In particular, I would like to express my gratitude to Amanda Lo for the productive collaboration and the delightful company throughout this research. Thanks to John Murphy and Rita Shen for the valuable discussions had within this research. I am also grateful to our industry partner, more specifically Matthew Tyler and John Taranto for all their advices and feedbacks on the prototype system.

Finally, my gratitude to my family and friends who made this work possible with their support and encouragement.

Contents

Nomenclature	xv
1 Introduction	1
1.1 Nature of the online accommodation industry	4
1.2 Research Scope	8
1.2.1 Ontology and Information Integration	9
1.2.2 Information Integration Process	10
1.2.3 Information Integration Architecture	11
1.3 Research Contribution	12
1.4 Research Methodology	15
1.4.1 Design Science Research in Information Systems	15
1.4.2 Apply the Design Science Research Methodology	16
1.5 Thesis Outline	19
2 Literature Review	21
2.1 Data Warehousing Approach	23
2.1.1 Database	23
2.1.2 Web Information	24
2.1.3 XML Data	25
2.1.4 ATDW	27
2.1.5 Conclusion	27

2.2	Wrapper-Mediator Approach	28
2.2.1	Centralized Integration Approach	29
2.2.1.1	Global-as-View (GaV)	30
2.2.1.2	Local-as-View (LaV)	36
2.2.1.3	Hybrid Approach	41
2.2.2	Distributed P2P Approach	42
2.3	Discussion	44
2.3.1	Scalability and Flexibility	45
2.3.2	Efficiency	46
2.3.3	Accuracy	48
2.3.4	Conclusion	50
2.4	Related Work on Semantic Web	52
2.4.1	E-Tourism Ontology	53
2.4.2	Semantic Web Query	55
2.4.3	Community based Query	57
3	A Semantic Integration Framework	59
3.1	Combining the existing integration approaches	60
3.1.1	Ontology and Data Heterogeneity	61
3.1.2	Integration Process	62
3.1.3	Integration Architecture	63
3.2	Overview of the semantic integration framework	64
3.2.1	Semantic Interoperability	67
3.2.2	Publish-Combine-Use Integration Process	68
3.2.3	Hybrid Integration Architecture	69
3.3	Ontology and Semantic Interoperability	69
3.3.1	Semantic Layer	71
3.3.2	Ontology	72

3.3.2.1	Data Model	73
3.3.2.2	Schema Mapping	76
3.3.2.3	Data Mediation	78
3.4	Publish-Combine-Use Integration Process	79
3.4.1	Publish	80
3.4.2	Combine	82
3.4.3	Use	83
3.5	Hybrid Integration Architecture and Query Processing	84
3.6	Conclusion	86
4	Ontology and Semantic Interoperability	88
4.1	Semantic Layer	89
4.2	Local-to-Global Schema Mapping	91
4.2.1	XML Schema and Ontology Comparison	92
4.2.2	Use of Ontology for Concept Mapping	97
4.2.2.1	Ontology Mediated XML Transformation	101
4.3	Global-to-Global Schema Mapping	106
4.3.1	Many-to-Many Ontology Mapping	107
4.3.1.1	Ontology Parsing	109
4.3.1.2	Classification Tree Construction	109
4.3.1.3	Concept Classification	110
4.3.2	Classification Tree	111
4.3.2.1	Granularity Calculation	112
4.3.2.2	Semantic Similarity Calculation	114
4.3.3	Concept Classification	115
4.3.4	Related Work	117
4.4	Conclusion	118

5	Information Integration Process	119
5.1	Publish-Combine-Use	120
5.1.1	Integration Example - Hotel Rate Sharing	122
5.1.2	Benefits	124
5.2	Information Group	124
5.3	Query Interface	125
5.3.1	Parameter Selection	128
5.3.2	XML Query Interface	130
5.3.3	Display Query Interface	132
5.4	Conclusion	133
 6	 Query Processing	 135
6.1	Query Layer Overview	136
6.2	Query Language	139
6.2.1	SPARQL	140
6.2.2	Triple Pattern	142
6.2.3	Global Triple	143
6.2.4	Local Triple	143
6.3	Query Message Generation	145
6.3.1	Graph Formation	145
6.3.2	Message Generation	148
6.4	Query Processing	149
6.4.1	Peer Selection	149
6.4.2	Query Resolving	150
6.4.3	Neighborhood Selection	156
6.4.4	Query Translation	157
6.5	Conclusion	159

7 Accommodation Hub - An Accommodation Information Integration Prototype System	161
7.1 Introduction	162
7.2 Accommodation Hub	164
7.2.1 Mapping Module	167
7.2.2 Query Module	168
7.2.3 Interface Module	169
7.3 Case Study	173
7.3.1 Hotel Rate	173
7.3.2 Hotel Review	175
7.4 Conclusion	177
8 Evaluation	178
8.1 Evaluation Design	178
8.1.1 Evaluation Measurements	180
8.1.2 Data Sets	185
8.1.3 Test Cases	187
8.2 Evaluation Results	190
8.3 Discussion and Conclusion	196
9 Conclusions	199
9.1 Future Research Directions	201
A Data Samples	203
A.1 Sample Schemas	203
A.2 Sample XML Documents	207
B Accommodation Hub	214
B.1 Provider Portal	214
B.1.1 Account Registration	214

CONTENTS

B.1.2	Log In	216
B.1.3	Provider Portal	218
B.1.4	Publish Information Source	218
B.1.5	Manage Source	220
B.2	Operator Portal	222
B.2.1	Source Management	223
B.2.2	Query Definition	224
B.2.3	Query Management	225
B.2.4	Ontology Management	227
B.3	User Portal	228
	References	242

List of Figures

1.1	The accommodation industry hierarchy	8
3.1	The semantic integration framework	66
3.2	Schema mapping and semantic interoperability	70
3.3	Hotel domain ontology	77
3.4	ontology mediated XML transformation	79
3.5	Information integration process	80
4.1	Semantic layer	89
4.2	Partial description of concept mapping ontology	98
4.3	Concept classification based mapping process	108
4.4	Classification tree	113
5.1	The use of query interface	126
5.2	Query interface editor	129
5.3	Graphical web form	134
6.1	Query processing	137
6.2	Hotel ontology graph	141
6.3	Ontology sub-graphs	144
6.4	Bottom-up query resolving	152
6.5	Hotel rate schema example	152

LIST OF FIGURES

6.6	CMO instances	153
6.7	Sample flow for $T_{x \rightarrow o}$	155
6.8	Sample hotel ontology	158
7.1	Accommodation hub	165
7.2	Mapping module	167
7.3	Query module	168
7.4	Interface module	170
7.5	Schema mapping editor	171
7.6	Query interfaces in user portal	172
8.1	Incorrect precision calculation	181
8.2	Total concept comparison cycle	191
8.3	XML to Ontology Transformation Time ($T_{transform}$)	193
8.4	Sample test results	194
8.5	Test results on increasing number of information sources	195
B.1	Accommodation hub	215
B.2	Account registration	216
B.3	Login form	217
B.4	Provider portal	218
B.5	Publish source	219
B.6	Mapping definition	220
B.7	Manage source	221
B.8	Operator portal	222
B.9	Source management	223
B.10	Query definition step 1	224
B.11	Query definition step 2	225
B.12	Query Management	226

LIST OF FIGURES

B.13 Ontology management	227
B.14 User portal	228

List of Tables

2.1	Integration approaches comparison	50
3.1	Data model comparison	74
4.1	XML schema and ontology difference	93
4.2	Conventional translation approaches	94
6.1	Concept mapping between homogeneous ontologies	158
8.1	Partial test results for ontology mapping evaluation	190
8.2	Partial test results for ontology mapping evaluation	191
8.3	Partial results for mapping generation	192